A Machine Learning Approach for Disease Surveillance and Visualization using Twitter Data

Ashwin Ashok, Guruprasad M, Prakash C O, Shylaja S S
Department of Computer Science and Engineering
PES University
Bangalore, India
ashwinashok1998@gmail.com, mguru1998@gmail.com, coprakasha@pes.edu, shylaja.sharath@pes.edu

***Abstract*—Insights from real-time disease surveillance systems are very useful for the public to take preventive measures against the diseases and it also benefits the pharmaceutical manufacturers in improving the sales of medicines for the particular disease and ensuring adequate availability of medicines when they are needed.**

A disease outbreak is an event wherein there is a rise in the number of positive cases for a disease in a short span of time. An outbreak can be limited to a particular region or time of the year. Diseases can be detected by several approaches, social media being preferred method due to availability of real-time data. Hence, data from social media, especially Twitter can be used to detect live events and monitor them efficiently. In order to detect diseases precisely, this paper proposes an approach wherein tweets, which are collected and pre-processed, can be effectively vectorized and clustered into the appropriate diseases with the use Agglomerative Clustering technique. The tweets can also be visualized using their geo information in order to generate zones which have high density of diseases. Such a surveillance system can be of use for early prediction of disease outbreaks, in turn facilitating faster and better handling of the situation.

Keywords—Twitter, Disease Surveillance, Natural Language Processing, Clustering, Visualization

I. INTRODUCTION

An event where there is an increase in the occurrences of a disease in a region is termed as a disease outbreak. Using the tweets posted by people, outbreaks can be effectively detected, so that measures can be taken by the people to protect themselves. Also, health organizations can undertake necessary measures to control the outbreak. The authors believe that the study would benefit budding researchers in the fields of data analysis and visualization, and also people in the health-care sector who would want to see the enormous value of Twitter data in providing beneficial insights during surveillance of diseases.

Twitter is an online social media platform where a lot of information is exchanged, mostly as messages called 'tweets' that offer insights into many aspects of life. Here, users share links, pictures, comments on live events and their experiences and opinions on a variety of topics. Twitter produces large amounts of data at an unprecedented scale of nearly 500 million tweets per day from close to 336 million active users. It is the availability of large amounts of twitter data, the large outreach of tweets and the ease with which it can be fetched that has made Twitter a very popular source for researchers looking for data to draw insights from. This makes the publicly available data a valuable resource for discovery of interesting healthcare insights. Hence, Twitter has also drawn the attention of organizations working towards improving the overall health of citizens, besides creating awareness resulting in detecting and preventing the spread of certain diseases. The authors feel that, compared to conventional methods for detection of diseases, analysis of data obtained from twitter is faster, economical and precise. The collected twitter data is cleaned, vectorized and passed through supervised or unsupervised learning algorithms to gain insights. Few such algorithms which are suited to twitter data include K-means clustering and Agglomerative clustering.

Agglomerative clustering, a type of hierarchical clustering, makes use of bottom-up approach in order to group data points into clusters. In this method, initially each data point is treated as a cluster. Iteratively, the closest pair of clusters are combined into one cluster and this process is repeated until a required number of clusters has been reached.

II. RELATED WORK

Twitter data has been widely used by researchers in the past in order to gain insights and develop models on various topics. Healthcare has been one of the leading domains of research due to the ease of availability of large amounts of data and algorithms, hence resulting in a large number of ways of solving a problem. In this section, a summary of the related works carried out in this domain in the past is presented.

Kathy Lee, Ankit Agarwal and Alok Choudhary [1] have described how spatial, temporal and textual analytics on Twitter data can be used to develop a scalable surveillance system for close monitoring of flu and cancer. Their work effectively conveys the Flu activity in the form of an interactive US Disease Surveillance Map, distribution of the various symptoms and treatments for flu and cancer and also a timeline of volume of flu and cancer related tweets using various charts. Their system is useful for early detection of rise in number of cases of diseases like flu and for observing patients suffering from different cancer types and the preferred treatments that would be taken.

An approach for reliable classification of tweets using influenza-based keywords was presented by Kenny Byrd, Alisher Mansurov and Olga Baysal [2]. In their work, they have presented how the growth in the number of cases of influenza in a region can be predicted accurately and also shown that the spread of influenza can be monitored in a few cities closely using a web-based mapping program.

Neha Garg and Rinkle Rani [3] have presented an approach for extracting Twitter data, preprocessing and geographically clustering them (clustering the tweets based on geographical co-ordinates) using K-means clustering.

978-1-5386-9471-8/19/$31.00 ©2019 IEEE

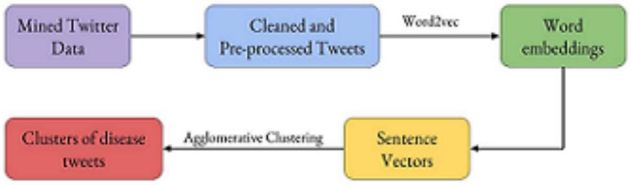

Fig. 1. Proposed Framework

They have also demonstrated how the Elbow Method can be used in order to generate the optimal number of clusters. Their work illustrates the usage of word cloud in order to visualize the most frequently occurring words.

Using a combination of deep learning and non-linear regression approaches, Bin Zou et al [4] have shown that Infectious Intestinal Diseases (IID) can be detected effectively and its extent measured with the aid of cues present in Twitter data. A set of keywords related to IID has been obtained and an IID vocabulary has been generated using Word2Vec, which is a deep learning model. The geolocation of tweets has been captured by obtaining either the exact co-ordinates of the user when he posts the tweet or by obtaining it from the profile of the user.

The previous works focus on using Twitter data for surveillance of a particular disease - either flu, influenza, IID or cancer, applying supervised learning techniques to classify tweets based on the sentiment and visualizing them on the map and geographical clustering of tweets using the K-means technique.

III. METHODOLOGY

Tweets were collected and cleaned, so as to retain only the required content. A list of keywords (here, diseases) was set up, based on which tweets were classified and labelled. Word vectors were generated, using which the Sentence vectors were generated. The tweets were segregated based on month, after which the count was obtained month-wise and the result was plotted. These tweets were also grouped into clusters of different diseases using different clustering techniques and visualized on a map. Data collection, pre-processing, generation of word embeddings, clustering and tweet analysis were carried out in Python using its libraries, APIs and packages.

A. Data Collection

Twitter data was collected from January 2018 to June 2018 using the REST API[1] which is used to interact with Twitter services. Tweepy, the library interface for the Twitter API provided access to the entire Twitter RESTful API methods. Each method accepted various parameters and returned responses. The Streaming API was used to retrieve tweets in real-time or to create a live feed using a user stream.

StreamListener, a method used by Tweepy to classify the most relevant tweets and directs them to suitably named methods, but these methods are only substitutes for larger methods. Stream, StdOutListener and Filter are the other methods used in the data collection process. The final dataset after applying filters consists of tweets which are that of diseases alone. From the dataset of disease tweets collected, there were a large number of retweets. These retweets were

{'created_at': 'Fri Jul 21 05:24:06 +0000 2017', 'id': 888268325062451203, 'text': '60 Chipotle customers sickened, 1 norovirus confirmed from Va. location https://t.co/tLgrwIRWkc', 'user': {'id': 81913437, 'name': 'Washington Press', 'screen_name': 'WashingtonCP', 'location': 'Washington, DC', 'geo_enabled': True, 'lang': 'en'}, 'geo': {'type': 'Point', 'coordinates': [38.90181532, -77.03733586]}, 'coordinates': {'type': 'Point', 'coordinates': [-77.03733586, 38.90181532]}, 'place': {'id': '01fbe706f872cb32', 'place_type': 'city', 'name': 'Washington', 'full_name': 'Washington, DC', 'country_code': 'US', 'country': 'United States', 'bounding_box': {'type': 'Polygon', 'coordinates': [[[-77.119401, 38.801826], [-77.119401, 38.99538], [-76.909396, 38.99538], [-76.909396, 38.801826]]]}, 'retweet_count': 0, 'favorite_count': 0, 'entities': {'hashtags': [], 'user_mentions': [], 'symbols': []}, 'favorited': False, 'retweeted': False}

Fig. 2. Attributes of a sample tweet considered for analysis

removed by only retaining tweets with original text in the dataset. The tweets which did not have any context to any of the diseases were also removed.

The tweets were stored in a file of JSON type in a semi structured format. The downloaded tweets contained attributes like date and time of creation, tweet ID, text contents, geo information data wherever it was available, favourite count, retweet count, details of the user who posted the tweet and the details of users who were mentioned in the tweet. The data is organized in key-value pair format, where key represents the tweet's attribute and value represents the value corresponding to that attribute (Fig. 2).

B. Pre-processing

Pre-processing is a challenging task in Natural Language Processing. The data collected is raw and contains undesired information. In order to eliminate this for an efficient analysis, certain pre-processing tasks were carried out. For the analysis, only English tweets were considered. Firstly, the contents of each tweet were split into individual words with whitespace acting as the delimiter. The tokenized words were then converted to lower case in order to maintain uniformity. Retweets and repeat instances of the original tweet were not considered in order to prevent redundancy. Stop words are words which impart little meaning to a sentence. Since they do not alter the context of a tweet, they were removed.

Certain tweets also contained unwanted symbols and emoticons which did not weigh much information and hence they too were removed. Modern messaging conventions has led to presence of a large number of slangs in the tweets, which were converted to synonyms for more efficient processing. As tweets have a character limit, it is obvious that abbreviations would be present, which were converted to their full forms in order to ensure that the information was captured clearly and fully. (Eg: KL to Kerala, Flu to Influenza)

Abbreviations were converted to full forms using a standard library. The standard library contains full forms of organizations, places, diseases, common abbreviations, word form of numbers and their abbreviated forms. Since abbreviations and full form of some Indian places were not present, they were manually added to a local copy of the library. During preprocessing, when an abbreviation was encountered, it was replaced by its respective expanded form by substituting it from the local copy, since the local copy had only one expanded form for each abbreviation. A similar library was used for dealing with slangs. Whenever slangs were encountered, they were dealt with in the same manner as abbreviations.

[1]https://developer.twitter.com/overview/documentation

978-1-5386-9471-8/19/$31.00 ©2019 IEEE

C. Embeddings

A computer cannot understand textual data, hence a numerical representation should be given for the words in the text. The basic way to represent these words in a computer is using a rule-based system like Word-Net which carries some issues, one of them being scalability where new words need to be manually added to the corpus. Another issue is use of standard similarity metrics (using the distance measures) due to lack of numbers. A word embedding is a learned representation of text where words having the same meaning in a particular context have similar numerical representations. The purpose of converting words to word embeddings is to have a form understandable by computers and hence can be used to train machine learning models.

Word2vec by Mikolov et al [5] is a model which was used to generate the embeddings for the words in the vocabulary. It takes into account the context of the word in the entire corpus by taking a fixed number of neighbors into consideration. Other models to generate word embeddings include term frequency-inverse document frequency(TF-IDF) which captures the weight of the word based on the number of occurrences of the word in the document rather than the context of the word in the document, bag-of-words (BOW) model, which is a simple representation that disregards grammar and word order and more particularly used for computer vision and n-grams, which is a sequence of n items of a sample of the text which faces an issue when it comes to out-of-vocabulary (OOV) words. For generating the embeddings, capturing the context of a word is important, hence Word2Vec was chosen.

1) Word-to-vec Conversion: The tweets, which are collected and preprocessed, are added into a vocabulary. The 50-dimensional vectors only for those words which appear atleast thirty times in the vocabulary, are trained using Word2vec using the gensim[2] model. The context window size five has been considered for these vectors which implies that the context words of a word are five words to the left and to the right of the word. Word2vec predicts the word from its context words (called the CBOW model) or predicts the context words using the word (Skip-gram model). With 50 dimensions, there is a tradeoff between processing time and effective representation. On increasing the dimension, there were no significant improvements while on lowering the dimension these embeddings could not effectively capture the semantics of the sentence hence resulting in poor performance of the model.

One word context is referred to as the CBOW model which tends to predict the probability of the word given a context. Multi-word context has no major differences from the one word context, except the type of probability distribution obtained and the type of hidden layer present. It is used to predict multinomial distribution given context of many words and to store details about the relation of the target word to other words from the corpus. Skip-gram model can be used to predict c related words having a target word for the input. This case is the opposite of CBOW multi-word model.

2) Sentence-to-vector generation: Tweet text cannot be expressed using Word2vec alone. Shylaja et al [6], in their work, studied the disadvantages of using standard text

Fig. 3. Attributes of a sample tweet considered for analysis

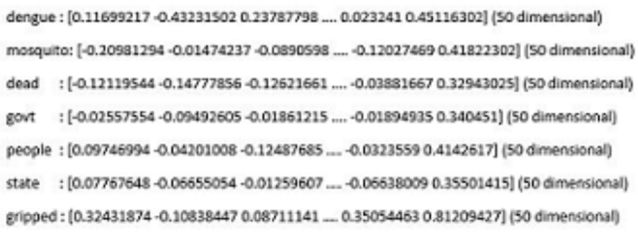

Fig. 4. Vectors for the words of the sample tweet

feature extraction techniques with no notion of context in vectorizing sentences and proposed the usage of sentence vectors which considered the importance of context of words occurring in a sentence. Since the task at hand also is largely dependent on context, inspired by various research experiments, the usage of sentence vectors was proposed by the aggregation of word embeddings of indiviual words in a sentence. For each word of the tweet in the vocabulary, the vector of that word was considered. The mean of all the vectors considered (Dimension-wise mean) was obtained, resulting in a vector, which is termed as the Sentence vector.

The tweet (in Fig. 3) was tokenized as: {“keraladengue350dead”, “kerala”, “left”, “govt”, “fails”, “people”, “stands”, “mosquitoes”, “state”, “gripped”, “dengue”}

From the list of tokens, “dengue”, “mosquito”, “dead”, “govt”, “people”, “state”, “gripped” were present in the vocabulary, hence their word vectors (Fig. 4) were considered and a 50-dimensional sentence vector was generated, which represented the tweet as a whole. The sentence vector generated for the sample tweet is: [0.03712477 -0.12952958 -0.00662458 ……. 0.01385842 0.44580534] (50 dim).

D. Tweet Analysis

The collected tweets were pre-processed in order to remove the stopwords, abbreviations, slangs, symbols, emoticons, retweets and duplicate tweets. Each tweet was labelled with the corresponding disease by considering the context and sentiment elucidated by it, following which the remaining tokenized contents, the date of posting and the label for every tweet were recorded. The tweet count for each disease was obtained by aggregating the number of tweets for each label.

A season consists of two or three months. Hence, tweets were grouped at month level in order to capture the season-wise distribution of various diseases and analysis of the same in order to determine the season-wise occurrences of various diseases. The preprocessed text contents were analyzed on six different diseases and a month wise disease count was depicted in a clustered column chart.

[2]https://radimrehurek.com/genism/models/word2vec.html

978-1-5386-9471-8/19/$31.00 ©2019 IEEE

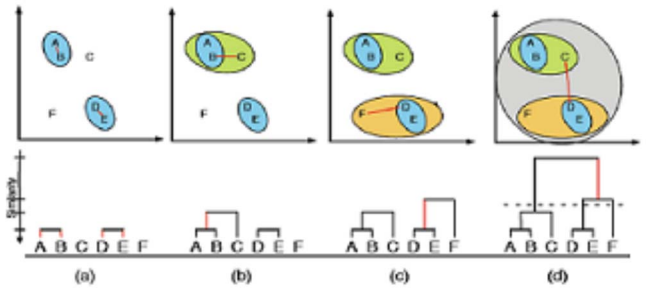

Fig. 5. Hierarchical Agglomerative Clustering

E. Clustering

The technique of identifying similar data points (here, tweets) in a data set and grouping them into various clusters is called clustering[3]. Entities in each cluster are comparatively more similar to entities of that cluster than those of the other clusters. The tweets were clustered into the various disease groups by the application of Agglomerative, K-means and Spectral clustering techniques on the sentence vectors and the results tabulated.

Agglomerative Clustering

In the usage of hierarchical clustering, the number of clusters need not be known beforehand. It allows us to specify any distance metric for calculating distance between points (where as in k-means, the distance metric is restricted to only Euclidean distance). Even if the points are scattered in a non-spherical manner, hierarchical clustering groups them into a single cluster. In case of spectral clustering, the algorithm works well only if points are clustered in a spherical manner.

The process of Hierarchical Agglomerative Clustering (Fig. 5) [11] applied is as follows:

1. Treat each tweet as a separate cluster, hence initial number of clusters is equal to the number of pre-processed tweets considered.
2. Construct a distance matrix for the clusters by calculating the Euclidean distance between each pair of clusters.
3. Among the various pair of clusters, the pair with the least distance is considered. Let the closest pair of clusters be 'r' and 's'.
4. The two clusters are then combined into a single cluster, hence reducing the total number of clusters by one. The corresponding row in the distance matrix is deleted and a new row corresponding to the new cluster is added.
5. The distances for the newly added row are computed iteratively as d[k,(r,s)] = min (d[k,r],d[k,s]) where k represents all other clusters taken one at a time.

The process is repeated until the number of rows left in the distance matrix is same as the number of required clusters, which is six. The number of clusters was chosen to be six since it is known that there are six diseases (clusters) which were considered and used to filter the collected tweets.

[3]http://scikit-learn.org/stable/modules/clustering.html

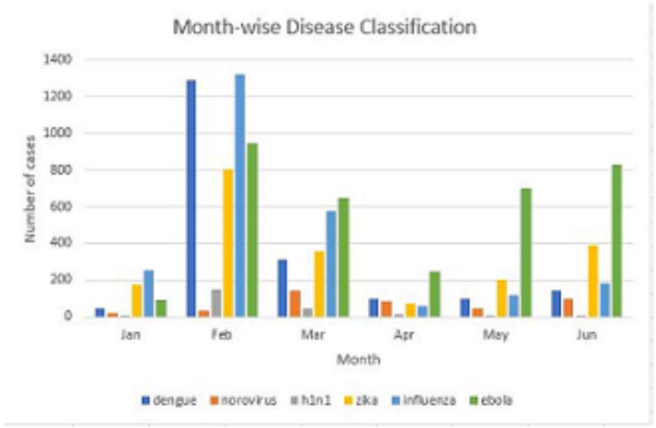

Fig. 6. Monthwise disease classification and count

TABLE I. COUNT OF TWEETS (IN NUMBERS) AFTER CLUSTERING

Clustering Technique	dengue	noro-virus	H1N1	zika	influ-enza	ebola
Actual	540	279	103	671	1089	964
Agglomerative	476	285	102	636	1208	798
K-Means	456	425	126	653	991	854
Spectral	135	126	31	591	1664	958

From Table I, it can be noted that Agglomerative, K-Means and Spectral clustering techniques have been experimented with, in which tweets are grouped into the respective clusters which mimic the diseases. Each clustering technique grouped the tweets into six clusters, among which Agglomerative clustering technique produced results which was closest to the actual count of the tweets of a particular disease.

IV. VISUALIZATION

Twitter allows users to add location information to their tweets. This location information, which is stored as a string, can be the GPS location of their device from which they post the tweet or the location of a network they are connected to. Twitter also gives an option to the users to tag a particular place as the location to their tweet. Due to privacy reasons, many users do not share the location, hence most tweets do not contain location information leaving behind around two percent of total tweets in the dataset which were tagged with the location.

The technique implemented is as follows:

- Each pre-processed tweet containing location information was considered. The latitudinal and longitudinal co-ordinates, along with the label for that tweet, were obtained. The entries for all such tweets containing location information were stored in a file.
- Using PHP, the geo co-ordinates and the label for each tweet were retrieved and further accessed for plotting using JavaScript, as the Google Maps API can be used through JavaScript, while PHP is needed for file operations.

978-1-5386-9471-8/19/$31.00 ©2019 IEEE

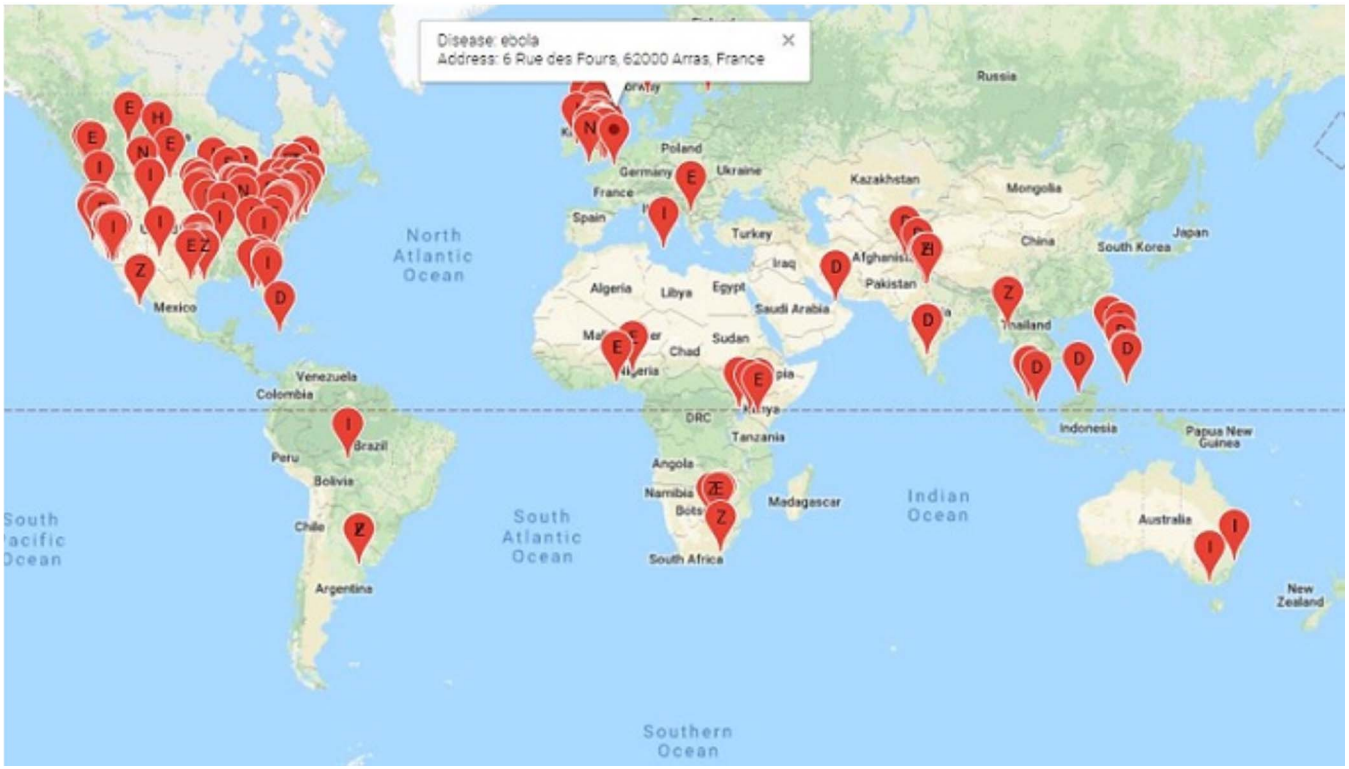

Fig. 7. Map depicting distribution of various diseases

- Using the API, a marker was laid at each of the co-ordinates on the Google Maps[4] layer and a label was set for each marker, corresponding to the tweet label.
- Using the same API, an approximate address for each tweet was determined and this address was linked to the respective marker.
- A heatmap layer was also generated by passing the co-ordinates of all tweets and the corresponding disease name.

V. RESULTS AND DISCUSSIONS

In this paper, the authors have presented an approach of collecting and pre-processing tweets in order to analyze and obtain insights from them. The usage of Word2vec has also been proposed in order to generate sentence vectors, which are used to represent the tweets. Agglomerative, K-Means and Spectral clustering techniques have been used to cluster the tweets into different diseases and the results have been compared.

TABLE II. EVALUATION RESULTS

Clustering Technique	Cosine Similarity	RMSE
Agglomerative	0.992	88.5767
K-Means	0.992	92.1312
Spectral	0.937	297.1192

From Table II, it can be seen that, the values of cosine similarities obtained on using Agglomerative and K-means clustering are identical, but on using a second evaluation metric, the RMSE, it can be seen that Agglomerative clustering gives a lower RMSE value compared to K-Means Clustering and also Spectral Clustering. Hence, it can be concluded that Agglomerative clustering performs better than both K-Means Clustering and Spectral Clustering. It was also observed that the obtained RMSE values were slightly high due to overlapping of clusters, which was caused by tweets depicting more than one disease.

From the clustered column chart (Fig. 6), it can be inferred that, influenza is present throughout the year, there is a large number of dengue cases in February but not as much as the other 5 months. There are few norovirus and H1N1 cases around, which is a good sign as both of them are deadly diseases. As with zika and ebola, a large number of cases can be observed in June compared to other months. The high volume of diseases in months like February and June could be due to the fact that seasonal changes occur in these months.

For each tweet described in Section IV, a marker has been overlaid on the map layer through a marker constructor of the map object. On clicking on a particular marker, the corresponding disease as well as the approximate address of the place is displayed (Fig. 7). From the map, it can be observed that the points are skewed in certain regions only. This is because, only a small number of tweets had the geolocation enabled, hence only an approximate representation could be obtained.

[4]https://developers.google.com/maps/documentation

978-1-5386-9471-8/19/$31.00 ©2019 IEEE

Fig. 8. Heatmap

A heatmap is a type of data visualization in which colors are used to represent the magnitude or density of the data values on a map. A heatmap layer has been created using Google Maps. The heatmap gives an approximate idea in real-time as to which regions are highly affected by diseases (Fig. 8). It can be inferred that regions near North-Eastern USA and Argentina are highly affected by diseases due to red colour on the heatmap.

VI. CONCLUSIONS AND FUTURE WORK

The highlight of this work, compared to the existing literature, is that the authors have elucidated the complete procedure, right from mining of the data till the presentation of the results on a custom map to monitor cases of the diseases. Text analysis was also performed to monitor occurrences of the diseases and month-wise counts have been depicted in a clustered column chart. Moreover, an unsupervised learning technique (Agglomerative clustering) has been used to cluster tweets more precisely into various diseases and a heatmap has been generated depicting the intensity of the diseases across the world.

The authors have proposed the usage of Word2vec in generating word vectors for tweets. Other methods of generating word vectors could be used in order to generate efficient word vectors. Similarly, a lower dimension of vectors could be used in order to train the supervised or unsupervised learning models to obtain similar or higher accuracy. Research can be carried out in order to identify methods of generating sentence vectors which represent a tweet better. Other clustering methods could also be explored in order to cluster the tweets more effectively. In addition to this, the presented model could also be used for clustering not only disease-based tweets but also for generic disaster classification or to cluster tweets of different topics.

REFERENCES

[1] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: demonstration on flu and cancer," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013, pp. 1474–1477.

[2] K. Byrd, A. Mansurov, and O. Baysal, "Mining twitter data for influenza detection and surveillance," in Proceedings of the International Workshop on Software Engineering in Healthcare Systems. ACM, 2016, pp. 43–49.

[3] N. Garg and R. Rani, "Analysis and visualization of twitter data using k-means clustering," in Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on. IEEE, 2017, pp. 670–675.

[4] B. Zou, V. Lampos, R. Gorton, and I. J. Cox, "On infectious intestinal disease surveillance using social media content," in Proceedings of the 6th International Conference on Digital Health Conference. ACM, 2016, pp. 157–161.

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, ser. NIPS'13. USA: Curran Associates Inc., 2013, pp. 3111–3119. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999792.2999959

[6] S. S. S, A. Narayanan, A. Venugopal, and A. Prasad, "Document embedding generation for cyber-aggressive comment detection using supervised machine learning approach," in Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017). Kolkata, India: NLP Association of India, December 2017, pp. 348–355. [Online]. Available: http://www.aclweb.org/anthology/W/W17/W17-7543

[7] A. Sechelea, T. Do Huu, E. Zimos, and N. Deligiannis, "Twitter data clustering and visualization." in ICT, 2016, pp. 1–5.

[8] Y. Hu, S. Farnham, and K. Talamadupula, "Predicting user engagement on twitter with real-world events." ICWSM, vol. 15, pp. 168–177, 2015.

[9] C. C. Aggarwal and K. Subbian, "Event detection in social streams," in Proceedings of the 2012 SIAM international conference on data mining. SIAM, 2012, pp. 624–635.

[10] F. Atefeh and W. Khreich, "A survey of techniques for event detection in twitter," Computational Intelligence, vol. 31, no. 1, pp. 132–164, 2015.

[11] "Hierarchical clustering technique." [Online]. Available: "http://machinelearningstories.blogspot.com/2017/09/hierarchical-clustering-bottom-up.html"

978-1-5386-9471-8/19/$31.00 ©2019 IEEE